AD-A104 861 RICE UNIV HOUSTON TX F/G 12/1
KARL PEARSON WAS RIGHT,(U)
FEB 78 D W SCOTT, R A TAPIA, J R THOMPSON E(40-1)-5046
UNCLASSIFIED NL

NATIONAL BUREAU OF STANDARDS SPECIAL PUBLICATION 503
Proceedings of Computer Science and Statistics: Tenth Annual Symposium on the Interface
Held at Nat'l. Bur. of Stds., Gaithersburg, MD, April 14-15, 1977. (Issued February 1978)

(15) EC(49-1)-146, PHS-NIH-17269 (11) Feb 78

(10) (6) KARL PEARSON WAS RIGHT,

David W. Scott, Baylor College of Medicine, Houston, Texas 77030
Richard A. Tapia and James R. Thompson
Rice University, Houston, Texas 77001

(12) 15

ABSTRACT

A discussion is made of nonparametric versus parametric methods for the estimation of probability densities. A new algorithm for nonparametric density estimation is given and its performance compared with state-of-the-art kernel estimation algorithms.

Key words: computational feasibility, maximum likelihood, Pearson family, kernel estimates, penalized maximum likelihood.

1. INTRODUCTION

Two major causes for poor (especially nonrobust) optimization theoretic techniques in statistics are

(1) an inappropriate choice of a parameter (function) space

and

(2) an inappropriate choice of a criterion function (functional).

"Appropriateness" is determined by a balance between computational feasibility and approximation to truth. It is to be expected that the advent of the high speed digital computer should drastically raise our pain threshold of computational feasibility. Consequently it is somewhat surprising that most standard statistical procedures have remained unchanged since the 1930's. Many of these involve the estimation of probability densities.

2. DISCUSSION

In 1922 Fisher [1] presented the concept of parametric maximum likelihood estimation. We recall that his development requires the functional form of the unknown density $f(x|\theta)$ be known. Given a random sample $\{x_1, x_2, \ldots, x_n\}$ from f, we seek that value $\hat{\theta}_n(\underline{x})$ contained in appropriate parameter space $\Theta \subset R$ which maximizes

$$\log f_n(\underline{x}|\theta) = \sum_{j=1}^{n} \log f(x_j|\theta) . \tag{1}$$

Then under very general conditions,

$$\hat{\theta}_n \xrightarrow{a.s.} \theta_o \tag{2}$$

and

$$\hat{\theta}_n \xrightarrow{d} N\left(\theta_o, \frac{-1}{nE\left(\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2}\right)}\right) . \tag{3}$$

The latter result is particularly appealing, since it states that the parametric maximum likelihood estimator asymptotically achieves the Cauchy-Schwarz (Cramer-Rao) lower bound for $E[(\check{\theta} - \theta)^2]$, where $\check{\theta} \in \check{\Theta}$, the class of unbiased estimates for θ.

The optimality properties of parametric maximum likelihood algorithms are likely to be of little utility if (as is generally the case) we do not have a good idea as to the functional form of the unknown density. For example, if we assume the density is normal, the maximum likelihood estimator for the median θ_{me} is $\bar{x}$. If, in fact, the underlying distribution is Cauchy, $\bar{x}$ is no better an estimator for θ_{me} than any single one of the observations. In general, if we assume an incorrect functional form of the density and use any of the classical parametric techniques for estimating the density, we will find that

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} E\left(f(x)_{est,n} - f(x)_{true}\right)^2 dx > 0 \ . \tag{4}$$

The pathology of parametric maximum likelihood estimation under real world conditions should not be unexpected. An optimization-theoretic technique designed to have good performance under very restrictive conditions (e.g., that the functional form of the density is known) is unlikely to perform well when we step outside the domain of these conditions. We need to devise algorithms which are "optimal" in a more general and realistic setting. This point was implicitly raised a quarter century before maximum likelihood by Karl Pearson [7]. (For a discussion of the Fisher-Pearson battle on maximum likelihood, the reader is referred to [13].) He considered a fairly large class of probability densities characterized by the differential equation

$$\frac{d \log f(x)}{dx} = \frac{x - a}{b_o + b_1 x + b_2 x^2} \ . \tag{5}$$

The estimation of the four parameters is readily carried out via the first four sample moments. Unfortunately, although the Pearson Family contains many of the classical distributions, it has serious deficiencies. For example, it contains no multimodal densities.

In order to obtain a practical extension of Pearson's concept to density estimation in the general setting where we know only that the underlying density is "smooth", we must develop an estimator where the number of characterizing parameters increases with the sample size. The simple histogram (dating back to John Graunt in 1662 [3]) has such a property but suffers from discontinuities. These may be eliminated quite readily by connecting midpoints with straight lines. The extreme "locality" of the histogram is less easily ameliorated.

Computationally more complicated but possessing better consistency properties than the histogram is the kernel density estimator (or "shifted histogram" [12], [6], [8]). Here, on the basis of a random sample $(x_1, x_2, \ldots, x_n)$ we have the estimator

$$\hat{f}_H(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right) \tag{6}$$

where K is any probability density having

$$\int_{-\infty}^{\infty} |K(y)| dy < \infty \tag{7}$$

$$\sup_{-\infty < y < \infty} |K(y)| < \infty \tag{8}$$

$$\lim_{y\to\infty} |yK(y)| = 0 \ . \tag{9}$$

To minimize the asymptotic integrated mean square error, we have the optimal

$$h = \left[\frac{9}{2\int (f''(x))^2 dx}\right]^{1/5} n^{-1/5}, \tag{10}$$

which gives as asymptotic integrated mean square error

$$IMSE = 2^{4/5}9^{-1/5}\,\frac{5}{4}\left[\int (f''(x))^2 dx\right]^{1/5} n^{-4/5} \tag{11}$$

Unfortunately, the design parameter h requires approximate knowledge of $\int (f''(x))^2 dx$. An iterative algorithm for the estimation of h is given in [12]. Monte Carlo results indicate that a twofold overestimation or underestimation of h typically causes a twofold increase of the IMSE over that shown in (11). A survey of other nonparametric density estimation techniques is given in [13].

A new approach motivated by a suggestion of Good [2] has been considered in [4], [5], [11], [13]. Here we seek that density $f \in H_o^s(a,b)$ which maximizes the criterion functional

$$L(f) = \sum_{j=1}^{n} \log f(x_j) - \sum_{k=0}^{s} \alpha_k \int_a^b (f^{(k)})^2 dx, \tag{12}$$

i.e.,

$$f^{(k)} \in L^2(a,b); \quad k = 0,1,\ldots,s$$

$$f^{(k)}(a) = f^{(k)}(b) = 0; \quad k = 0,1,2,\ldots,s-1$$

$$f \geq 0$$

$$\int_a^b f(x)dx = 1 .$$

The solution to (12) is referred to as the maximum penalized likelihood estimator. From [5] we have

Theorem. The MPLE estimator exists and is unique. ∎

Recently, a discretized approximation to the solution of (12) has been algorithmitized and investigated by Scott [10], [11]. This work suggests

Theorem. If $\hat{f}_n(\cdot)$ is the solution to the MPLE criterion and $f_T \in H_o^s(a,b)$ then

$$\int_a^b E[(\hat{f}_n(x) - f_T(x))^2]dx \xrightarrow{n} 0 \tag{13}$$

where $f_T(\cdot)$ is the density f truncated to (a,b). ∎

From a practical standpoint, the performance of $\hat{f}_n(\cdot)$ is relatively insensitive to the selection of the design parameters $\underset{\sim}{\alpha}$. If we set all the $\alpha_j = 0$ except for α_2, it is not unusual for a change of α_2 by a factor of 100 from the optimal to increase the IMSE by less than a factor of 2.

In Table 1, we compare the IMSE of the MPLE with that of popular Gaussian kernel estimator for various densities and sample sizes. Of special note is the fact that although we have used the optimal (and unobtainable) design parameter for the kernel estimator, we have used the suboptimal value of $\alpha_2 = 10$ throughout for the MPLE estimator.

TABLE 1

IMSE Values of the MPLE ($\alpha_2 = 10$) and Gaussian Kernel Density Estimation (with optimal h) for Various Distributions and Sample Sizes.

Density	n	MPLE IMSE	Kernel IMSE
N(0,1)	25	.0027	.0041
	100	.00079	.00129
	400	.00033	.00053
$\frac{1}{2}$N(-1.5,1)	25	.00159	.00128
$+\frac{1}{2}$N(1.5, 1)	100	.00054	.00052
t_5	25	.00282	.00475
	100	.00084	.00157

3. CONCLUSIONS

The supposed optimality of classical parametric density estimation procedures is frequently invalid because the true functional form of the density is unknown. Nevertheless, we can attack the more general and practical problem of estimating a density of unknown functional form. The maximum penalized likelihood density estimator has been algorithmitized and is now a part of standard statistical software [11].

4. ACKNOWLEDGEMENTS

The authors wish to thank the U.S. Office of Naval Research, the U.S. Air Force Office of Scientific Research, the U.S. Energy Research and Development Administration and the National Heart, Lung and Blood Institute for their support of this work under grants NR042-283, AFOSR76-2711, E-(40-1)-5046 and NIH 17269 respectively.

5. BIBLIOGRAPHY

[1] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London, Series A, 222, 309-368.

[2] Good, I.J. (1971). Nonparametric roughness penalties for probability densities. Biometrika, 58, 255-277.

[3] Graunt, John (1662). Natural and Political Observations on the Bills of Mortality.

[4] de Montricher, G.M. (1973). Nonparametric Bayesian Estimation of Probability Densities by Function Space Techniques, Doctoral dissertation, Rice University, Houston, Texas.

[5] de Montricher, G.M., Tapia, R.A., and Thompson, J.R. (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods. Annals of Statistics, 3, 1329-1348.

[6] Parzen, Emmanuel (1962). On estimation of a probability density function and mode. Annals of Mathematical Statistics, 33, 1065-1076.

[7] Pearson, Karl (1895). Contributions to the mathematical theory of evolution. II. Skew variations in homogeneous material. Philosophical Transactions of the Royal Society of London, Series A, 186, 343-414.

[8] Rosenblatt, Murray (1956). Remarks on some nonparametric estimates of a density function. Annals of Mathematical Statistics, 27, 832-835.

[9] Schoenberg, I.J. (1972). Notes on spline functions II on the smoothing of histograms. MRC Technical Report 1222.

[10] Scott, D.W. (1976). Nonparametric Probability Density Estimation by Optimization Techniques. *Doctoral dissertation*, Rice University, Houston, Texas.

[11] Scott, D.W. (1977). A software package for the nonparametric estimation of probability densities (subroutine NDMPLE). International Mathematical and Statistical Libraries.

[12] Scott, D.W., Tapia, R.A. and Thompson, J.R. (1977). Kernel density estimation revisited. Nonlinear Analysis, 1, 339-372.

[13] Tapia, R.A. and Thompson, J.R. (1977). Nonparametric Probability Density Estimation. The Johns Hopkins University Press.